# The development of an information criterion for Change-Point Analysis.

Paul A. Wiggins and Colin H. LaMont
*Departments of Physics, Bioengineering and Microbiology, University of Washington, Box 351560.*
*3910 15th Avenue Northeast, Seattle, WA 98195, USA[*]*



Change-point analysis is a flexible and computationally tractable tool for the analysis of times series data from systems that transition between discrete states and whose observables are corrupted by noise. The change-point algorithm is used to identify the time indices (change points) at which the system transitions between these discrete states. We present a unified information-based approach to testing for the existence of change points. This new approach reconciles two previously disparate approaches to Change-Point Analysis (frequentist and information-based) for testing transitions between states. The resulting method is statistically principled, parameter and prior free and widely applicable to a wide range of change-point problems.

## I. INTRODUCTION

The problem of determining the true state of a system that transitions between discrete states and whose observables are corrupted by noise is a canonical problem in statistics with a long history (e.g. [1]). The approach we discuss in this paper is called Change-Point Analysis and was first proposed by E. S. Page in the mid 1950s [2, 3]. Since its inception, Change-Point Analysis has been used in a great number of contexts and is regularly re-invented in fields ranging from geology to biophysics [1, 4, 5].

The primary goal of this paper is to develop a new information-based approach to Change-Point Analysis which simplifies its application in problems, including those where a specific change-point statistics have not been computed. We approach Change-Point Analysis from the perspective of *Model Selection* and *Information Theory*. Akaike pioneered a powerful approach to Model Selection by the minimization of the Kullback-Leibler Divergence [6], a measure of information loss by approximating the true process with a model [7, 8]. He demonstrated that two key principles of modeling, predictivity and parsimony, were in fact conceptually and mathematically linked (e.g. [8]). In short, the addition of superfluous parameters to a model, reducing parsimony, results in information loss, reducing predictivity (e.g. [8]). Akaike derived an unbiased estimator for information loss, the Akaike Information Criterion (AIC), which proved to be at once exceptionally tractable and widely applicable.

Unfortunately Akaike's approach is limited to regular models [9]. Change-Point Analysis and many other applications are singular. These models contain unidentifiable parameters with nearly zero Fisher Information, which greatly increase the complexity of the model and lead to the catastrophic failure of AIC to estimate information loss. The subject of this paper is the implementation of information-based model selection in the context of Change-Point Analysis. We have recently proposed a Frequentist Information Criterion (FIC) applicable even in the context of singular models. Using FIC and an approximation analogous to that used by Akaike to derive AIC, we develop a model criterion that accounts for the unidentifiability of the change-point indices. Importantly, this criterion does not depend on the detailed form of the model for the individual states but only on the number of model parameters, in close analogy with AIC. Therefore we expect this result to be widely applicable anywhere the change-point algorithm is applied.

Frequentist statistical tests have already been defined for a number of canonical change-point problems. It is therefore interesting to examen the relation between this approach and our newly-derived information-based approach. We find the approaches are fundamentally related. The information-based approach can be understood to provide an predictively-optimal confidence level for a generalized ratio test. The Bayesian Information Criterion (BIC) has also been used in the context of Change-Point Analysis. We find very significant differences between our results and the BIC complexity that suggest that BIC is not suitable for application to change-point analysis since it can lead to either over or under segmentation of the data, depending on the specific context.

---

[*]Electronic address: pwiggins@uw.edu; URL: http://mtshasta.phys.washington.edu/

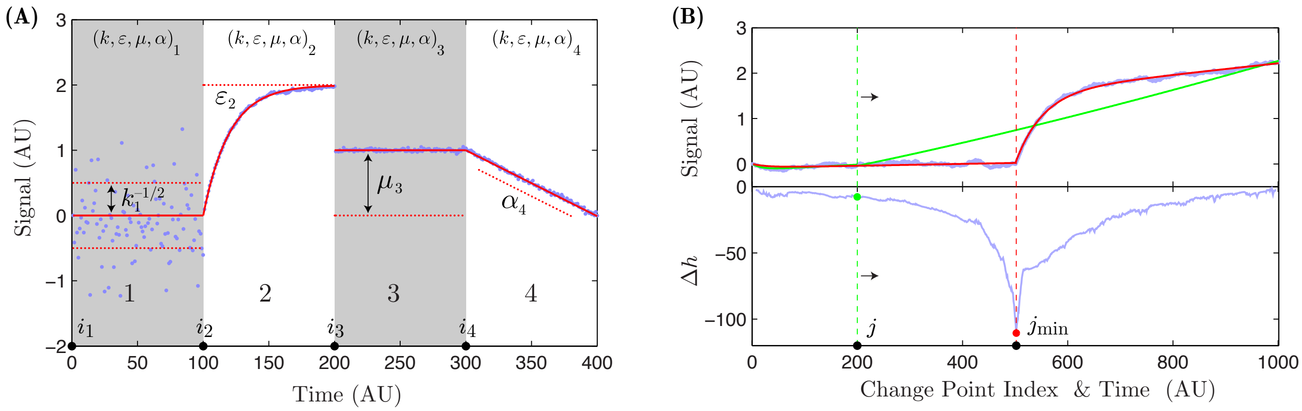


FIG. 1: **Panel A: State model schematic.** A state model for biophysical applications is parameterized by four model parameters that are written as the vector $\boldsymbol{\theta} \equiv (k, \varepsilon, \mu, \alpha)$. Above we schematically illustrate the role of each parameter in shaping the signal. **Panel B: Schematic of binary segmentation.** To segment a partition, the information change due to placing a change point at each time index along the x axis is considered. (The dashed red and green lines represent possible change points.) For each change-point index, a minimum-information fit is performed on the two resulting data partitions (top panel, blue dots), resulting in the solid curves (top panel, red and green for the respective change-point indices). For each change-point index, an information change is computed (bottom panel). The change point is placed at the time index that minimizes the information change (red dashed line).

## II. PRELIMINARIES

We introduce the following notation for a signal: a set of $N$ ordered observations from a one-dimensional stochastic process[1]:

$$X^N \equiv (X_1, X_2, ..., X_N) \sim p(\cdot), \tag{1}$$

where the observation index is often but not exclusively temporal and the probability distribution for the stochastic process is represented as $p$. We shall represent the probability distribution for the model $\mathscr{M}$ as:

$$q(X^N|\mathscr{M}), \tag{2}$$

where these is no guarantee that true distribution is a member of the model family.

**Information and cross entropy.** The coding information for signal $X^N$ given model $\mathscr{M}$ is:

$$h(X^N|\mathscr{M}) \equiv -\log q(X^N|\mathscr{M}), \tag{3}$$

and the cross entropy for the signal (average coding information) is:

$$H^N(\mathscr{M}) \equiv \underset{p(\cdot)}{\mathbb{E}_X} h(X^N|\mathscr{M}), \tag{4}$$

where the expectation over the signal $X^N$ is understood to be taken over the true distribution $p$.

**The Change-Point Model.** We define a model for the signal corresponding to a system transitioning between a set of discrete states. We define the discrete time index corresponding to the start of the $I$th state $i_I$. This index is called a *change point*. The model parameters describing the signal in the $I$th interval are $\boldsymbol{\theta}_I$. Together these two sets of parameters ($i_I$ and $\boldsymbol{\theta}_I$) parameterize the model . The model parameterization for the signal (including multiple states) can then be written explicitly:

$$\mathscr{M}^n = \begin{pmatrix} 1 & i_2 & \dots & i_n \\ \boldsymbol{\theta}_1 & \boldsymbol{\theta}_2 & \dots & \boldsymbol{\theta}_n \end{pmatrix}, \tag{5}$$

---

[1] When $X$ appears in upper case, it should be understood as a random variable whereas it is a normal variable when it appears in lower case. If we need a statistically independent set of variables of equal size, we will use the random variables $Y^N$, which have identical properties to the $X^N$.

where $n$ is the number of states or change points. A schematic example of a change-point mode for a biophysical signal is shown in Figure 1. The two sets of parameters ($\boldsymbol{\theta}_I$ and $i_I$) are fundamentally different. We shall assume that the state model is regular: i.e. the parameters $\boldsymbol{\theta}_I$ have non-zero Fisher information [10]. By contrast, the change-point indices $i_I$ are discrete and typically non-harmonic parameters. For instance, consider a true model $p = q$ where $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. In this scenario the cross entropy will be independent of $i_2$ as long as $i_2 \in (i_1, i_3)$. The Fisher information corresponding to $i_2$ is threfore zero. These properties have important consequences for model selection [10].

**Determination of model parameters.** Fitting the change-point model is performed in two coupled steps. Given a set of change-point indices $\boldsymbol{i}^n \equiv (i_1, ..., i_n)$, the maximum likelihood estimators (MLE) of the state model parameters $\boldsymbol{\Theta}^n \equiv (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n)$ are defined:

$$\hat{\boldsymbol{\Theta}}^n_X = \arg\min_{\boldsymbol{\Theta}^n} h(X^N|\mathscr{M}^n). \tag{6}$$

The determination of the change-point indices $\boldsymbol{i}^n$ is a nontrivial problem since not only are the change-point indices unknown, but even the number of transitions ($n$) is unknown.

**Binary Segmentation Algorithm.** To determine the change-point indices, we will use a binary-segmentation algorithm that has been the subject of extensive study (e.g see the references in [4]). In the global algorithm, we initialize the algorithm with a single change point $i_1 = 1$. The data is sequentially divided into partitions by binary segmentation. Every segmentation is *greedy*: i.e. we choose the change point on the interval $(1, N)$ that minimizes the information in that given step, without any guarantee that this is the optimum choice over multiple segmentations. The family of models generated by successive rounds of segmentation are said to be *nested* since successive changes points are added without altering the time indices of existing change points. Therefore, the previous model is always a special case of the new model. The binary segmentation process is shown schematically in Fig. 1, Panel B. In each step, after the optimum index for segmentation is identified, we statistically test the change in information (due to segmentation) to determine whether the new states are statistically supported. The change-point determined by binary segmentation determine the change-points in the MLE model $\hat{\mathscr{M}}^n$. The local binary-segmentation algorithm differs from the global algorithm only in that we consider the binary segmentation of each partition of the data independently. The algorithms as described explicitly in the supplement.

**Information-based model selection.** The model that minimizes the cross entropy (Eqn. 4) is the most predictive model. Unfortunately, the cross entropy cannot be computed since the expectation cannot be taken with respect to the true but unknown probability distribution $p$ in Eqn. 4. The natural estimator of the cross entropy is the information (Eqn. 3), but this estimator is biased from below: Due to the phenomena of over-fitting, added model parameters always reduce the information (or equivalently the training error) even as the predictivity of the model is reduced by the addition of superfluous parameters. We must therefore construct an unbiased estimator of the cross entropy which we call the *information criterion*:

$$\mathrm{IC}(X^N, n) \equiv h(X^N|\hat{\mathscr{M}}^n_X) + \mathscr{K}(n), \tag{7}$$

where $\mathscr{K}$ is the complexity of the model which is defined as the bias in the information as an estimator of cross-entropy:

$$\mathscr{K}(n) \equiv \mathbb{E}_{\substack{X,Y\\p}} \left\{ h(Y^N|\hat{\mathscr{M}}^n_X) - h(X^N|\hat{\mathscr{M}}^n_X) \right\}, \tag{8}$$

where the expectations are taken with respect to the true distribution $p$ and $X^N$ and $Y^N$ are independent signals. For a regular model in the asymptotic limit, the complexity is equal to the number of model parameters and the information criterion is equal to AIC. In the context of singular models, a more generally applicable approach must be used to approximate the complexity.

**Frequentist Information Criterion.** The Frequentist Information Criterion (FIC) uses a more general approximation to estimate the model compleixty. Since the true distribution $p$ is unknown, we make a frequentist approximation, computing the complexity for the model $\mathscr{M}$ as a function of the true parameterization:

$$\mathscr{K}_{\mathrm{FIC}}(\mathscr{M}^n, n) \equiv \mathbb{E}_{\substack{X,Y\\q(\cdot|\mathscr{M}^n)}} \left\{ h(Y^N|\hat{\mathscr{M}}^n_X) - h(X^N|\hat{\mathscr{M}}^n_X) \right\}, \tag{9}$$

and the corresponding information criterion is defined:

$$\mathrm{FIC}(X^N, n) \equiv h(X^N|\hat{\mathscr{M}}^n_X) + \mathscr{K}_{\mathrm{FIC}}(\hat{\mathscr{M}}^n_X, n), \tag{10}$$

where the complexity is evaluated at the MLE parameters $\hat{\mathscr{M}}_X^n$. The model that minimizes FIC has the smallest expected cross entropy.

**Approximating the FIC complexity.** The direct computation of the FIC complexity (Eqn. 9) appears daunting, but a tractable approximation allows the complexity to be estimated. The complexity difference between the models is:

$$\mathscr{k}(n) \equiv \mathscr{K}_{\mathrm{FIC}}(n) - \mathscr{K}_{\mathrm{FIC}}(n-1), \tag{11}$$

which is called the nesting complexity. An approximate piecewise expression can be computed as follows. Let the observed change in the MLE information for the $n$th nesting be

$$\Delta h_n \equiv h(X^N|\hat{\mathscr{M}}_X^n) - h(X^N|\hat{\mathscr{M}}_X^{n-1}), \tag{12}$$

where $n$ denotes the $n$th nesting of model $\mathscr{M}$. Consider two limiting cases: When the new parameters are identifiable, let the nesting complexity be given by $\mathscr{k}_+$ whereas when the new parameters are unidentifiable, let the nesting complexity be given by $\mathscr{k}_-$. When the new parameters are identifiable, the model is essentially regular therefore:

$$\mathscr{k}_+ = \mathscr{d}, \tag{13}$$

where $\mathscr{d}$ is the number of harmonic[2] parameters added to the model in the nesting procedure, as predicted by AIC.

To compute $\mathscr{k}_-$, we assume the unnested model is the true model and compute the complexity difference in Eqn. 11. We then apply a piecewise approximation for evaluating the nesting complexity [10]:

$$\mathscr{k}(n) \approx \begin{cases} \mathscr{k}_-(n), & -\Delta h_n < \mathscr{k}_-(n) \\ \mathscr{k}_+(n), & \text{otherwise} \end{cases}. \tag{14}$$

Since the nesting complexity represents complexity differences, the complexity can be summed:

$$\mathscr{K}_{\mathrm{FIC}}(n) \equiv \sum_{j=1}^{n} \mathscr{k}(j), \tag{15}$$

where the first term in the series, $\mathscr{k}(1)$ is computed using the AIC expression for the complexity. An exact analytic description of the complexity remains an open question.

## III. AN INFORMATION CRITERION FOR CHAGNE-POINT ANALYSIS

**Complexity of a state model.** As a first step towards computing the complexity for the change-point algorithm, we will compute the complexity for a signal with only a single state. It will be useful to break the information into the information per observation. Using the Markov property of the process, the information associated with the $i$th observation is:

$$h_i(X^N|\boldsymbol{\theta}) \equiv -\log q(X_i|X_{i-1};\boldsymbol{\theta}). \tag{16}$$

For a stationary process, the average information per observation is constant $\overline{h} \equiv \mathbb{E}\, h$. The fluctuation in the information $\delta h_i \equiv h_i - \overline{h}$ has the property that it is independent for each observations:

$$\mathbb{E}\,\delta h_i\,\delta h_j = C_0 \delta_{ij}, \tag{17}$$

where $C_0$ is a constant and $\delta_{ij}$ is the Kronecker delta, due to the Markovian property. In close analogy to the derivation of AIC, we will Taylor expand the information in terms of the model parameterization $\boldsymbol{\theta}$ around the true parameterization $\boldsymbol{\theta}_0$. We make the following standard definitions:

$$\delta\boldsymbol{\theta} \equiv \boldsymbol{\theta} - \boldsymbol{\theta}_0, \tag{18}$$

$$\hat{\boldsymbol{I}}_i \equiv \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}^T h_i(X^N|\boldsymbol{\theta}_0), \tag{19}$$

$$\boldsymbol{I} \equiv \mathbb{E}_X \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}^T h_i(X^N|\boldsymbol{\theta}_0), \tag{20}$$

$$\boldsymbol{x}_i \equiv \nabla_{\boldsymbol{\theta}} h_i(X^N|\boldsymbol{\theta}_0), \tag{21}$$

$$\boldsymbol{X} \equiv \sum_i \boldsymbol{x}_i. \tag{22}$$

[2] Harmonic parameters are parameter with sufficiently large Fisher information that they are not unidentifiable.

where $\delta\boldsymbol{\theta}$ is the perturbation in the parameters, $\boldsymbol{I}$ and $\hat{\boldsymbol{I}}_i$ are the Fisher Information and its estimator respectively. The subscript $i$ refers to the $i$th observation. Note that since the true parameterization minimizes the information by definition, $\mathbb{E}\,\boldsymbol{x}_i = 0$. Furthermore, Eqn. 17 implies that

$$\mathbb{E}\,\boldsymbol{x}_i\boldsymbol{x}_j^T = \boldsymbol{I}\delta_{ij} \tag{23}$$

where $\boldsymbol{I}$ is the Fisher Information. The Taylor expansion of the information can then be written:

$$h(X^N|\boldsymbol{\theta}) \;=\; h(X^N|\boldsymbol{\theta}_0) + \delta\boldsymbol{\theta}^T\boldsymbol{X} + \tfrac{1}{2}\delta\boldsymbol{\theta}^T N\boldsymbol{I}\delta\boldsymbol{\theta} + \mathcal{O}(\delta\boldsymbol{\theta}^3), \tag{24}$$

to quadratic order in $\delta\boldsymbol{\theta}$.

It is convenient to transform the random variables $\boldsymbol{x}_i$ to a new basis in which the Fisher Information is the identity. This is accomplished by the transformation

$$\boldsymbol{x}'_i \;\equiv\; \boldsymbol{I}^{-1/2}\boldsymbol{x}_i, \tag{25}$$

$$\boldsymbol{\theta}' \;\equiv\; \boldsymbol{I}^{1/2}\boldsymbol{\theta}, \tag{26}$$

which results in the following expression for the information:

$$h(\boldsymbol{\theta}|X_I) \;=\; h(X^N|\boldsymbol{\theta}_0) + \delta\boldsymbol{\theta}'^T\boldsymbol{X}' + \tfrac{1}{2}N\delta\boldsymbol{\theta}'^T\delta\boldsymbol{\theta}' + \mathcal{O}(\delta\boldsymbol{\theta}^3). \tag{27}$$

In our rescaled coordinate system, $\boldsymbol{X}'$ can be interpreted as an unbiased random walk of $N$ steps with unit variance in each dimension.

We determine the MLE parameter values:

$$\delta\hat{\boldsymbol{\theta}}'_X = -\tfrac{1}{N}\boldsymbol{X}'. \tag{28}$$

To compute the complexity we need the following expectations of the information:

$$\mathbb{E}_{X,Y}\; h(Y^N|\hat{\boldsymbol{\theta}}_X) \;=\; \mathbb{E}_{X,Y}\;\left\{h(Y^N|\boldsymbol{\theta}_0) - \tfrac{1}{N}\boldsymbol{X}'^T\boldsymbol{Y}' + \tfrac{1}{2N}\boldsymbol{X}'^2 + \mathcal{O}(\delta\boldsymbol{\theta}^3)\right\}, \tag{29}$$

$$\mathbb{E}_{X}\; h(X^N|\hat{\boldsymbol{\theta}}_X) \;=\; \mathbb{E}_{X,Y}\;\left\{h(X^N|\boldsymbol{\theta}_0) - \tfrac{1}{2N}\boldsymbol{X}'^2 + \mathcal{O}(\delta\boldsymbol{\theta}^3),\right\}. \tag{30}$$

Since the signals $X^N$ and $Y^N$ are independent, the second term on the RHS of Eqn. 29 is exactly zero. It is straight forward to demonstrate that

$$\mathbb{E}_X\boldsymbol{X}_I'^2 = N\mathscr{d}, \tag{31}$$

where $\mathscr{d}$ is the dimension of the parameter $\boldsymbol{\theta}$, which has an intuitive interpretation as the mean squared displacement ($\boldsymbol{X}'^2$) of a unbiased random walk of $N$ steps in $\mathscr{d}$ dimensions. The complexity is therefore:

$$\mathcal{K} \equiv \mathbb{E}_{X,Y}\;\left\{h(Y^N|\hat{\boldsymbol{\theta}}_X) - h(X^N|\hat{\boldsymbol{\theta}}_X)\right\} = \mathscr{d}. \tag{32}$$

which is the AIC complexity. To compute the complexity associated with the first binary segmentation, we will compute the nesting complexity $\mathscr{k}(2)$ using Eqn. 14. We will therefore generate the observations $X^N$ and $Y^N$ using the unsegmented model $\mathcal{M}^1$. Remember that by convention we assign the first change-point index to the first observation $i_1 = 1$. The optimal but fictitious change-point index for binary segmentation is:

$$\hat{\imath}_2(X) \;=\; \arg\min_{1<i\le N}\,\{\; h(X^{[1,i-1]}|\hat{\boldsymbol{\theta}}_{X^{[1,i-1]}}) + h(X^{[i,N]}|\hat{\boldsymbol{\theta}}_{X^{[i,N]}})\;\}\,, \tag{33}$$

where the $X^{[j,k]}$ represent the respective partitions of the signal $X^N$ made by the change point $i$. (Note that in the case of an AR process, it is possible to write overlapping partitions to account for the system memory.) The MLE model for two states is defined:

$$\hat{\mathcal{M}}^2_X \;\equiv\; \begin{pmatrix} 1 & \hat{\imath}_2 \\ \hat{\boldsymbol{\theta}}_{X^{[1,\hat{\imath}_2-1]}} & \hat{\boldsymbol{\theta}}_{X^{[\hat{\imath}_2,N]}}) \end{pmatrix}. \tag{34}$$

To compute the nesting complexity, we compute the difference in the information between the two-state and one-state MLE models:

$$\begin{aligned} h(X^N|\hat{\mathcal{M}}^2_X) - h(X^N|\hat{\mathcal{M}}^1_X) \;=\; & \min_{1<i\le N}\,\{\; \cancel{h(X^{[1,i-1]}|\boldsymbol{\theta}_0)} + \cancel{h(X^{[i,N]}|\boldsymbol{\theta}_0)} - \cancel{h(X^{[1,N]}|\boldsymbol{\theta}_0)} \\ & -\tfrac{1}{2(i-1)}\boldsymbol{X}'^2_{[1,i-1]} - \tfrac{1}{2(N+1-i)}\boldsymbol{X}'^2_{[i,N]} + \tfrac{1}{2N}\boldsymbol{X}'^2_{[1,N]}\;\}\,, \end{aligned} \tag{35}$$

where the terms that are zeroth order in the perturbation cancel since the model is nested and $\boldsymbol{X}'_{[i,j]}$ are the $\boldsymbol{X}'$ computed in the two partitions of the data. (This equation is analogous to Eqn. 30.) It is straightforward to compute the analogous expression for information difference for signal $Y^N$. The nesting penalty can then be written:

$$\begin{aligned} \mathscr{k}_-(2) &\equiv \underset{q(\cdot|\mathscr{M}_0^1)}{\mathbb{E}_{X,Y}} \left\{ h(Y^N|\hat{\mathscr{M}}_X^2) - h(X^N|\hat{\mathscr{M}}_X^2) - h(Y^N|\hat{\mathscr{M}}_X^1) + h(X^N|\hat{\mathscr{M}}_X^1) \right\} \quad (36) \\ &= \underset{q(\cdot|\mathscr{M}_0^1)}{\mathbb{E}_X} \max_{1<i\le N} \left\{ \tfrac{1}{i-1}{\boldsymbol{X}'}^2_{[1,i-1]} + \tfrac{1}{N+1-i}{\boldsymbol{X}'}^2_{[i,N]} - \tfrac{1}{N}{\boldsymbol{X}'}^2_{[1,N]} \right\}, \quad (37) \end{aligned}$$

where the cross terms between signals $X^N$ and $Y^N$ are zero since the signals are independent. It is now convenient to introduce a $d$-dimensional discrete Brownian bridge:

$$\boldsymbol{B}'_j \equiv \boldsymbol{X}'_{[1,j]} - \tfrac{j}{N}\boldsymbol{X}'_{[1,N]}, \quad (38)$$

by using the well known relation between Brownian walks and bridges [11]. The Brownian bridge has the property that $\boldsymbol{B}'_0 = \boldsymbol{B}'_N = 0$, where each step has unit variance per dimension and mean zero. After some algebra, the nesting complexity can be written:

$$\mathscr{k}_-(2) = \underset{q(\cdot|\mathscr{M}_0^1)}{\mathbb{E}_X} \max_{1\le j<N} \left\{ \tfrac{N}{j(N-j)}{\boldsymbol{B}'}^2_j \right\}. \quad (39)$$

The details of the state model will determine the distribution function for the discrete steps in the Brownian bridge, but the Central Limit Theorem implies that the distribution will approach the normal distribution. Therefore, it is convenient to approximate the discrete Brownian bridge $\boldsymbol{B}'_n$ as an idealized Brownian bridge with normally distributed steps:

$$\boldsymbol{B}'_j \to \boldsymbol{B}_j \equiv \sum_{i=1}^{n} \boldsymbol{b}_i, \text{ such that } \boldsymbol{B}_N = 0, \quad (40)$$

where the $\boldsymbol{b}_i$ are steps that are normally distributed with variance one per dimension $d$ and mean zero. We now introduce a new random variable $U(N, d)$, the $d$-dimensional Change-Point Statistic [12, 13]:

$$U(N, d) \equiv \tfrac{1}{2} \max_{1\le j<N} \tfrac{N}{j(N-j)}\boldsymbol{B}_j^2, \quad (41)$$

which is a $d$-dimensional generalization of the change-point statistic computed by Hawkins [14]. In terms of the statistic $U$, the nesting penalty is

$$\mathscr{k}_-(2) = 2\,\mathbb{E}_U\, U(N, d) = 2\overline{U}(N, d). \quad (42)$$

We will discuss the connection to the frequentist LPT test shortly.

**Nesting complexity for $n$ states.** The generalization of the analysis to $n$ states is intuitive and straightforward. In the local binary-segmentation algorithm, segmentation is tested locally. The relevant complexity is computed with respect to the length of the $J$th partition. It is convenient to work with the approximation that all partitions are of equal length since the complexity is slowly varying in $N$. We therefore define the local nesting complexity

$$\mathscr{k}_{L-}(n) = 2\,\mathbb{E}_U\, U(\tfrac{N}{n-1}, d) = 2\overline{U}(\tfrac{N}{n-1}, d), \quad (43)$$

where $\frac{N}{n-1}$ is the mean partition length. The nesting complexity for the binary segmentation of a single state is show in Fig. 2 for several different dimensions $d$, and compared with the complexity predicted by AIC and BIC.

In the global binary-segmentation algorithm, the next change-point is chosen by identifying the best position over all intervals. We therefore generalize all our expressions accordingly. We introduce a generalization of the Change-Point Statistic where we replace $N$ with a vector of the lengths of the constituent segment lengths $\boldsymbol{N}^n \equiv (N_1, ...N_n)$. We now define our new change-point statistic:

$$U_G(\boldsymbol{N}^n, d) \equiv \max_{1\le i\le n} U(N_i, d). \quad (44)$$

Because it is computationally intensive to compute $U_G$ for all possible segmentations $\boldsymbol{N}^n$, we assume that all the partitions are roughly the same size and consider $n$ segments length $N/(n-1)$. Since the complexity is slowly

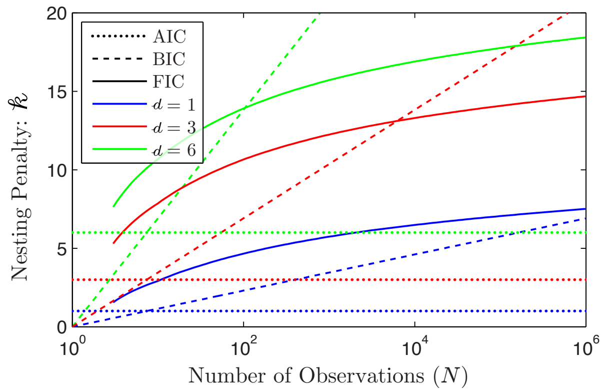


FIG. 2: **Nesting complexity for AIC, FIC and BIC.** The nesting complexity is plotted for three state dimensions $d = \{1, 3, 6\}$ and $n = 2$. First note that the AIC penalty is much smaller than the other two nesting complexities. BIC is empirically known to produce acceptable results under some circumstances. For sufficiently large samples ($N$), the $\mathscr{k}_{\rm BIC} > \mathscr{k}_{\rm FIC}$, resulting in over penalization and the rejection of states that are supported statistically. This effect is more pronounced for large state dimension $d$ where the crossover occurs for small observation number $N$. $\mathscr{k}_{\rm BIC}$ is too small for a wide range of sample sizes, resulting in over segmentation.

varying in $N$, this does not in general lead to significant information loss. We therefore introduce another change-point statistic:

$$\mathscr{k}_{G-}(n) \;\equiv\; 2\,\mathbb{E}_U \max_{1\le i\le n} \{\, U_i(\tfrac{N}{n-1}, d)\,\} \quad (\approx 2\,\mathbb{E}_U\, U_G(\boldsymbol{N}^n, d)\,) \tag{45}$$

that we will apply in the global binary-segmentation algorithm.

**Series expressions for the nesting complexity.** It is straightforward to compute the asymptotic dependence of the nesting penalty on the number of observations $N$:

$$k_{G-}(n) \;\approx\; 2\log\log\tfrac{N}{n} + 2\log n + d\log\log\log\tfrac{N}{n} + \ldots, \tag{46}$$
$$k_{L-}(n) \;\approx\; 2\log\log\tfrac{N}{n} + d\log\log\log\tfrac{N}{n} + \ldots \tag{47}$$

These expression are slowly converging and in practice, we advocate using Monte Carlo integration to determine the nesting penalty. If computationally cumbersome, Eqn. 46 and 47 are useful in placing our approach in relation to existing theory.

Both the local and the global encoding have the same leading-order $2\log\log N$ dependence that has been advocated by Hannan and Quinn [15], although interestingly not in this context. In contrast, this $2\log\log N$ dependence is in disagreement with the Bayesian Information Criterion, which has often been applied to change-point analysis. As illustrated by Fig. 2, the BIC complexity:

$$\mathscr{K}_{\rm BIC} = \tfrac{d}{2}\log N, \tag{48}$$

can be either too large or too small depending on the number of observations and the dimension of the model. It has long been appreciated that BIC can only be strictly justified in the large-observation-number limit. In this asymptotic limit, the BIC complexity is always larger than the FIC complexity due to the leading order $\log N$ dependence which will tend to lead to under fitting or under segmentation. It is clear from Fig. 2 that large ($N > 10^6$) may constitute much larger datasets than are produced in many applications.

**Global versus local complexity.** We proposed two possible parameter encoding algorithms above that give rise two distinct complexities: $k_{L-}$ and $k_{G-}$. Which complexity should be applied in the typical problem? For most applications, we expect the number of states $n$ to be proportional to the number of observations $N$. Doubling the length of the dataset will result in the observation of twice as many change points on average. The application of the local nesting complexity clearly has this desired property since it depends on the ratio of $N/n$. It is this complexity we advocate under most circumstances.

In contrast the global nesting complexity contains an extra contributions to the complexity $2\log n$. The reason is intuitive: In the global binary segmentation algorithm, one picks the best change point among $n$ segments and therefore complexity must reflect this added degree of choice. Consequently a larger feature must be observed

to be above the expected background. The use of the global nesting complexity makes a statement of statistical significance against the entire signal, not just against a local region. In the context of discussing the significance of the observation of a rare state that occurs just once in a dataset, the global nesting complexity is the most natural metric of significance.

**Computing the complexity from the nesting complexity.** To compute the FIC complexity, we sum the nesting complexities using Eqn. 15. For datasets with identifiable change points, the FIC complexity is initially identical to AIC:

$$\mathcal{K}_{\rm FIC}(n) = n d, \tag{49}$$

until the change in the information on nesting $\Delta h < \hbar_-$, when FIC predicts that there is a change in slope in the penalty. The FIC, AIC, and BIC predicted complexities are compared with the true complexity for an explicit change-point analysis in Fig. 3, Panel C. It is immediately clear from this example that FIC quantitatively captures the true dependence of the penalty, including the change in slope at $n = 4$, exactly as predicted by the FIC complexity. As predicted, the AIC complexity is initially correct until the segmentation process must be terminated. At this point the complexity increases significantly with the result that the AIC complexity fails to terminate the segmentation process. In contrast, the BIC complexity is initially too large, but fails to grow at a sufficient pace to match the true complexity for $n > 4$.

## IV. THE RELATION BETWEEN FREQUENTIST AND INFORMATION-BASED APPROACH

Consider the LPT test for the following problem: We propose the binary segmentation of a single partition. In the null hypothesis ($H_0$) is the partition is described by a single state (unknown model parameters $\boldsymbol{\theta}_0$) and the hypothesis to be tested ($H_1$) is that the partition is actually sub-divided into two states (unknown change point and model parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$). We use the log-likelihood ratio as the test statistic:

$$V(X^N) \equiv \log \frac{q(X^N|\hat{\mathcal{M}}_X^2)}{q(X^N|\hat{\mathcal{M}}_X^1)} = h(X^N|\hat{\mathcal{M}}_X^1) - h(X^N|\hat{\mathcal{M}}_X^2). \tag{50}$$

In the Neyman-Pearson approach to hypothesis testing, we assume the null hypothesis (1 state) and compute the distribution in the test statistic $V$. As before, we will expand the information around the true parameter values $\boldsymbol{\theta}_0$. In exact analogy to Eqn. 35, we find that $V$ and our previously defined statistic $U$ identically distributed:

$$V \sim U, \tag{51}$$

up to the approximations discussed in the derivation. Therefore we will simply refer to $V$ as $U$.

In the canonical frequentist approach we specify a critical test statistic value $u_\gamma$ above which the alternative hypothesis is accepted. $u_\gamma$ is selected such that the alternative hypothesis $H_1$ is rejected given that the null hypothesis $H_0$ is true with a probability equal to the confidence level $\gamma$:

$$\gamma = F_U(u_\gamma), \tag{52}$$

where $F_U$ is the cumulative distribution of $U$.

Therefore we can interpret both the information-based approach and the frequentist approach as making use of the same statistic $U$. In the frequentist approach, a confidence level ($\gamma$) is specified to determine the critical value $u_\gamma$ with which to accept the two-state hypothesis. The information-based approach also uses the statistic $U$, but the critical value of the statistic ($\hbar_-$) is computed from the distribution of the statistic itself $\hbar_- = 2\overline{U}$. The information-based approach chooses the confidence level that optimizes predictivity.

## V. APPLICATIONS

In the interest of brevity we have not included analysis of either experimental data or simulated data with a signal-model dimension larger than one, but we have tested the approach extensively. For instance, we have applied this technique to an experimental single-molecule biophysics application that is modeled by an Ornstein-Uhlenbeck process with state-model dimension of four [16]. We also applied the approach in other biophysical contexts including the analysis of bleaching curves, cell and molecular-motor motility [17].

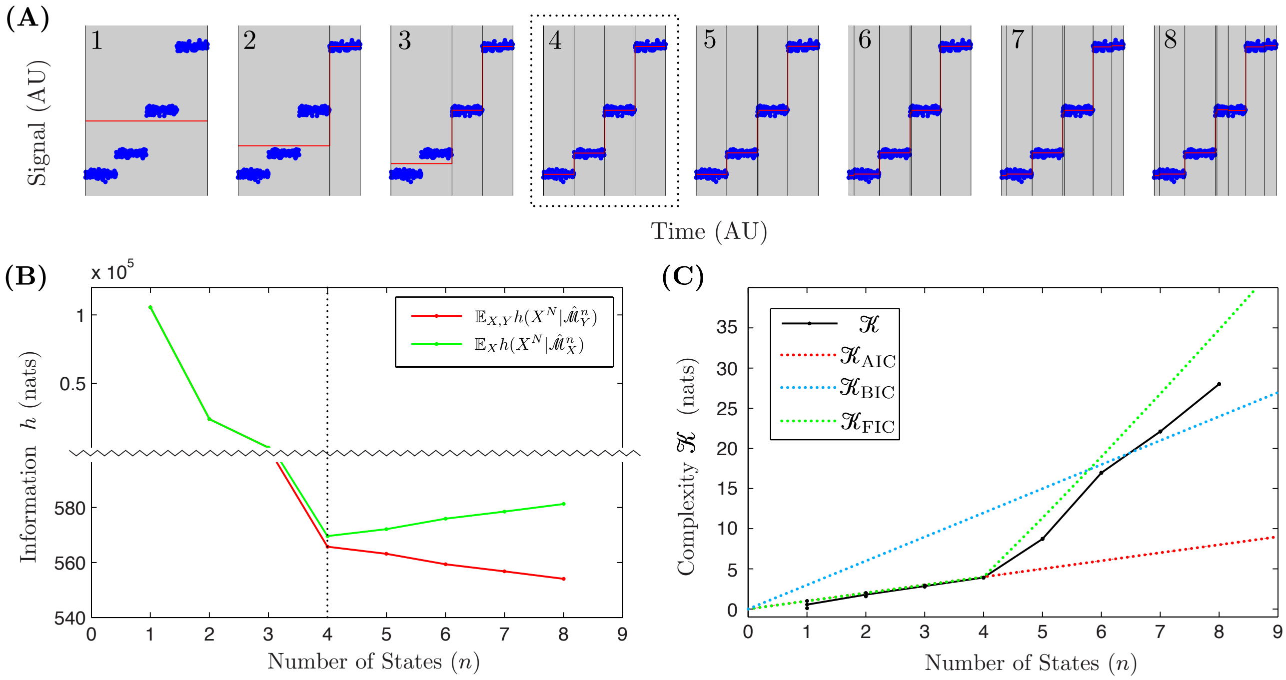


FIG. 3: **Information-based model selection. Panel A: Nested models generated by a Change-Point Algorithm.** Simulated data (blue points) generated by a true model with four states is fit to a family of nested models (red lines) using a Change-Point Algorithm. Models fit with $1 \le n \le 8$ states are plotted. The fit change points are represented as vertical black lines. The number of states ($n$) in each fit model is shown in the top-left corner of each panel. The true model has four states and the fit model with four states is indicated with a dotted box. The models with five through eight states have superfluous states that are not present in the true model. **Panel B: Four changes points minimizes information loss.** Both the expectation of the information (red) and the cross entropy (green) are plotted as a function of the number of states $n$. The y-axis ($h$, information) is split to show the initial large changes in $h$ as well as the subsequent smaller changes for $4 \le n \le 8$. The cross entropy (green) is minimized by the the model that best approximates the truth ($n = 4$). The addition of parameters leads to an increase in cross entropy (less predictive) as a consequence of the addition of superfluous parameters, as indicated by the increase of the cross entropy (green) for $n \ge 4$. The information loss estimator (red) is biased and continues to decreases with the addition of states as a consequence of over fitting. In an experimental context only the information can be computed since the true distribution is unknown. **Panel C: Complexity of Change-Point Analysis.** The true complexity is computed for the model shown in panel A via Monte Carlo simulation for $10^6$ realizations of the observations $X^N$ and compared with three models for the complexity AIC, FIC and BIC. For models with states numbering $1 \le n \le 4$, the true complexity (black) is correctly estimated by the AIC complexity (red dotted) and the FIC complexity (green). But for a larger number of states ($4 \le n \le 8$), only FIC accurately estimates the true complexity.

## VI. DISCUSSION

In this paper, we present an information-based approach to change-point analysis using the Frequentist Information Criterion (FIC). The information-based approach to inference provides a powerful framework in which models with different parameterization, including different model dimension, can be compared to determine the most predictive model. The model with the smallest information criterion has the best expected predictive performance against a new dataset.

Our approach has two advantages over existing frequentist-based ratio tests for change-point analysis: (i) We derive an FIC complexity that depends only on the dimension of the state model ($d$), the number of states ($n$) and observations ($N$). Therefore it may be unnecessary to develop and compute custom statistics for specific applications. (ii) In the frequentist approach one must specify an *ad hoc* confidence level to perform the analysis. In the information-based approach, the confidence level is chosen automatically based upon the model complexity. The information-based approach is therefore parameter and prior free.

As the number of change-points increases, the model complexity is observed to transition between an AIC-like complexity $\mathcal{O}(N^0)$ and a Hannan-and-Quinn-like complexity $\mathcal{O}(\log\log N)$. We propose an approximate piecewise expression for this transition. The computation of this approximate model complexity can be interpreted as the expectation of the extremum of a $d$-dimensional Brownian bridge. We believe this information-based approach to change-point analysis will be widely applicable.

### Author Contributions

P.A.W. and C.H.L. designed research; performed research; contributed analytic tools; analyzed data; or wrote the paper.

### Acknowledgments

P.A.W. and C.H.L. would like to thank K. Burnham, J. Wellner, L. Weihs and M. Drton for advice and discussions, D. Dunlap and L. Finzi for experimental data and M. Lindén and N. Kuwada for advice on the manuscript. This work was supported by NSF MCB grant 1243492.

---

[1] M. A. Little and N. S. Jones, Proc Math Phys Eng Sci **467**, 3088 (2011).
[2] E. S. Page, Biometrika **42**, 523 (1955).
[3] E. S. Page, Biometrika **44**, 248 (1957).
[4] J. Chen and A. K. Gupta, Communications in Statistics–Simulation and Computation **30**, 665 (2007).
[5] M. A. Little and N. S. Jones, Proc Math Phys Eng Sci **467**, 3115 (2011).
[6] S. Kullback and R. Leibler, Annals of Mathematical Statistics **22**, 79 (1951).
[7] H. Akaike, in *2nd International Symposium of Information Theory.*, edited by P. B. N. and E. Csaki (Akademiai Kiado, Budapest., 1973), pp. 267–281.
[8] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference.* (Springer-Verlag New York, Inc., 1998), 2nd ed.
[9] S. Watanabe, *Algerbraic geometry and statistical learning theory.* (Cambridge Univeristy Press, 2009).
[10] P. A. Wiggins, In preparation. (2015).
[11] Wikipedia, *Brownian bridge — wikipedia, the free encyclopedia* (2015), [Online; accessed 19-May-2015].
[12] L. Horváth, The Annals of Statistics **21**, 671 (1993).
[13] L. Horváth, P. Kokoszka, and J. Steinebach, Jounral of Multivariate Analysis **68**, 96 (1999).
[14] D. M. Hawkins, Journals of the american statistical association. **72**, 180 (1977).
[15] E. Hannan and B. G. Quinn, Journal of the Royal Statistical Society, Series B. **41**, 190 (1979).
[16] P. A. Wiggins, Submitted to Biophys J. (2015.).
[17] P. A. Wiggins, In preparation (2015.).
[18] A. Khinchine, Fundamenta Mathematica **6**, 9 (1924).
[19] A. Kolmogoroff, Mathematische Annalen **101**, 126 (1929).
[20] Wikipedia, *Law of the iterated logarithm — wikipedia, the free encyclopedia* (2015), [Online; accessed 19-May-2015].
[21] D. A. Darling and P. Erdös, Duke Math J. **23**, 143 (1956).
[22] Wikipedia, *Gumbel distribution — wikipedia, the free encyclopedia* (2015), [Online; accessed 19-May-2015].

**Global Binary-Segmentation Algorithm**

1. Initialize the change-point vector: $\boldsymbol{i} \leftarrow \{1\}$
2. Segment model $\hat{\mathcal{M}}(\boldsymbol{i})$:
   (a) Compute the entropy change that results from all possible new change-point indices $j$:
   $$\Delta h_j \leftarrow \hat{h}(\{i_1, ..., j, ..., i_n\}|X) - \hat{h}(\boldsymbol{i}|X), \tag{53}$$
   (b) Find the minimum information change $\Delta h_{\min}$, and the corresponding index $j_{\min}$.
   (c) **If** the information change plus the nesting complexity is less than zero:
   $$\Delta h_{\min} + \hbar_{G-} < 0 \tag{54}$$
   **then** accept the change-point $j_{\min}$
      i. Add the new change-point to the change-point vector.
      $$\boldsymbol{i} \leftarrow \{i_1, ..., j_{\min}, ...i_{n+1}\} \tag{55}$$
      ii. Segment model $\hat{\mathcal{M}}(\boldsymbol{i})$
   (d) **Else** terminate the segmentation process.

TABLE I: A global algorithm for binary segmentation. The information $\hat{h}$ is implicitly evaluated at the MLE state-model parameters $\hat{\boldsymbol{\Theta}}$.

#### 1. Type I errors (false positives)

In terms of the Cummulative Probability Distribution (CDF), the probability of a false positive change-point is:

$$\alpha = 1 - F_U(2\overline{U}), \tag{59}$$

where $U$ is the relevant change-point statistic and $\overline{U}$ is its expectation. Using the local binary-segmentation algorithm, $\alpha$ corresponds to the probability of a false positive per data partition and the change-point statistic is defined by Eqn. 41 evaluated at the average partition length $N_p \equiv \frac{N}{n}$. The false positive change-point acceptance probability is plotted in Figure 4.

The analogous false positive rate for the global binary-segmentation algorithm describes the probability of a false positive in the entire data set, including all partitions. In this cases, we use the change-point statistic defined by Eqn. **??**.

#### 2. Asymptotic form of the complexity function

In order to discuss the scaling of the complexity relative to the BIC complexity, we need to derive an asymptotic form for the complexity in the large $N$ limit. We do not recommend explicitly using this asymptotic expression for the complexity for Change-Point Analysis since it converges to the true complexity very slowly, especially for large $d$.

First let us consider related results for and Brownian walk rather than a Brownian bridge. Let us define $S_n$ as follows:

$$S_n \equiv |\boldsymbol{Z}_n| \tag{60}$$

$$\boldsymbol{Z}_n \equiv \sum_{i=1}^{n} \boldsymbol{z}_i \tag{61}$$

where the $\boldsymbol{z}_i$ are independent normally-distributed random variables with mean zero variance one per dimension

**Local Binary-Segmentation Algorithm**

1. Initialize the change-point vector: $\boldsymbol{i} \leftarrow \{1\}$, $I \leftarrow 1$.
2. Segment model $\hat{\mathcal{M}}(\boldsymbol{i})$ on state $I$:
   (a) Compute the entropy change that results from all possible new change-point indices $j$ on the interval $[i_I, i_{I+1})$:

$$\Delta h_j \leftarrow \hat{h}(\{...i_I, j, i_{I+1}, ...\}|X) - \hat{h}(\boldsymbol{i}|X), \tag{56}$$

   (b) Find the minimum information change $\Delta h_{\min}$, and the corresponding index $j_{\min}$.
   (c) **If** the information change plus the nesting complexity is less than zero:

$$\Delta h_{\min} + \hbar_{-\mathrm{L}} < 0 \tag{57}$$

   **then** accept the change-point $j_{\min}$
      i. Add the new change-point to the change-point vector.

$$\boldsymbol{i} \leftarrow \{..., i_I, j_{\min}, i_{I+1}, ...\} \tag{58}$$

      ii. Segment model $\hat{\mathcal{M}}(\boldsymbol{i})$ on states $I$ and $I+1$.
      iii. Merge the resulting index lists.
   (d) **Else** terminate the segmentation process.

TABLE II: A local algorithm for binary segmentation. The information $\hat{h}$ is implicitly evaluated at the MLE state model parameters $\hat{\boldsymbol{\Theta}}$.

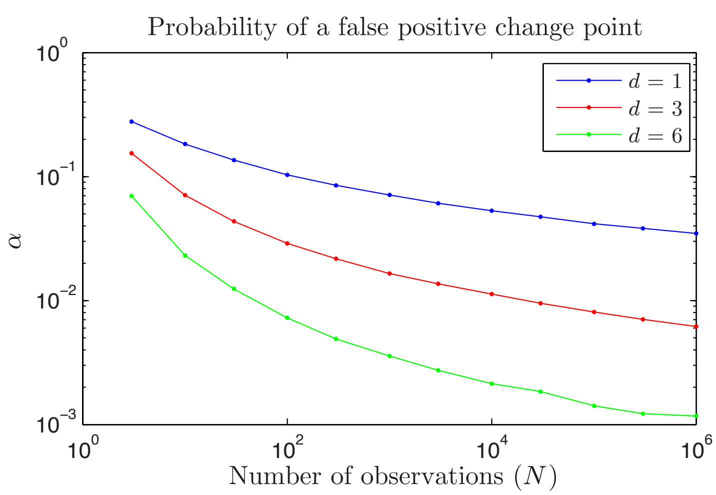


FIG. 4: **Probability of a false positive change-point.** The probability of a false positive change-point is shown as a function of the number of observations in the interval length $N$ for three different model dimensions.

$d$. The Law of Iterated Logs states that [18–20]:

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{n \log\log n}} = \sqrt{2} \quad \text{a.s.}, \tag{62}$$

where a.s. is the acronym for almost surely. (See Figure 5.) This behavior of $S_n$ is described in more detail by the Darling-Erdös Theorem [21]. Let us define a new random variable

$$U'(N_p) \equiv \max_{1\leq n\leq N_p} \frac{S_n}{\sqrt{n}}, \tag{63}$$

in $d = 1$ dimensions, the asymptotic cumulative distribution of $U'$ approaches the cumulative distribution for a Gumbel Distribution [21]:

$$\lim_{N_p \to \infty} \Pr\left[U' < \beta t + u\right] = \exp\left[-e^{-t}\right], \tag{64}$$

$$\beta(N_p) \equiv (2 \log\log N_p)^{-1/2}, \tag{65}$$

$$u(N_p) \equiv \beta\left[\beta^{-2} + \tfrac{1}{2}\log\log\log N_p - \log 2\pi^{1/2}\right], \tag{66}$$

where $\Pr$ denotes probability and the distribution parameters $u$ and $\beta$ are called the location and scale respectively and the average partition length is $N_p \equiv \frac{N}{n}$. Let us introduce the cumulative distribution function for $U$:

$$F_U(U) \equiv \Pr\left[U'_{(n)} < U\right]. \tag{67}$$

This expression can be reordered to put it in the canonical form of the Gumbel Distribution [22]:

$$F_U(U) = \exp\left[-\exp\left(-\tfrac{U-u}{\beta}\right)\right], \tag{68}$$

We can then use the well known expression in terms the cdf to compute the cdf of the maximum of $n$ random variables $U'$:

$$\Pr\left[U'_{(n)} < U_{(n)}\right] = F_U^n(U), \tag{69}$$

$$= \left(\exp\left[-\exp\left(-\tfrac{U-u}{\beta}\right)\right]\right)^n, \tag{70}$$

$$= \exp\left[-\exp\left(-\tfrac{U-u_n}{\beta}\right)\right], \tag{71}$$

where

$$u_n \equiv u + \beta \log n. \tag{72}$$

The mean and variance of the Gumbel Distribution are well known, allowing us to compute the expectation of $U'^2_{(n)}$:

$$\mathbb{E}_x\, U'^2_{(n)} \approx (u_n + \cancel{\beta\gamma})^2 + \cancel{\frac{\pi^2}{6}\beta^2}, \tag{73}$$

$$\approx 2\log\log N_p + 2\log n + \ldots \tag{74}$$

where $\gamma$ is the Euler-Mascheroni constant and we have used the cancel notation to show which terms have been dropped to lowest order. In the second line, we have written the expression to lowest order in $N$ and $n$.

Horváth has generalized the Darling-Erdös Theorem for a Brownian bridge in $d$ dimensions for the application to Change-Point Analysis in the context of the LPT test [12, 13]. The generalized expression for the cumulative distribution leads to a change in the expression for $u$ only:

$$u_d(N_p) \equiv \beta\left[\beta^{-2} + \tfrac{d}{2}\log\log\log N_p - \cancel{\log\Gamma\left(\tfrac{d}{2}\right)}\right] \tag{75}$$

where $\Gamma$ is the Gamma Function. We drop the last term since it is not leading order for large $N_p$. We now follow the same steps to generate the distribution for the maximum of $n$ random variables $U'$, leading to a new Gumbel Distribution with location $\mu_{n,d}$:

$$u_{n,d}(N_p) = \beta\left[\beta^{-2} + \tfrac{d}{2}\log\log\log N_p + \log n\right] \tag{76}$$

We now recompute the expectation for $d$ dimensions:

$$k_-^{\mathrm{G}}(N_p, n, d) \equiv \mathbb{E}_x\, U'^2_{(n)}(N_p, n, d), \tag{77}$$

$$\approx (u_{n,d} + \beta\gamma)^2 + \frac{\pi^2}{6}\beta^2 \tag{78}$$

$$\approx 2\log\log N_p + 2\log n + d\log\log\log N_p + \ldots \tag{79}$$

where we have kept terms only to highest order in $n$ and $N_p$.

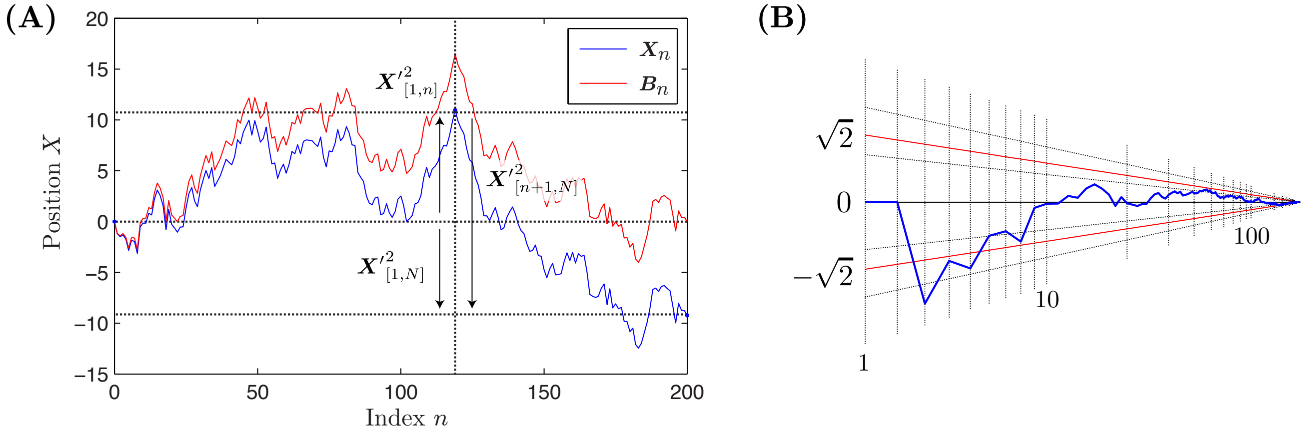


FIG. 5: **Panel A: Brownian Walk and Brownian Bridge.** A visualization of a random walk $\boldsymbol{X}'_{[}1, n]$ (blue) and the corresponding Brownian bridge $\boldsymbol{B}'_n$ (red). **Panel B: Law of Iterated Logs.** A visualization of $S_n/\sqrt{n \log\log n}$ (blue) plotted as an orthographic projection as a function of $n$. $\sqrt{2}$ (red) is the limit of the supremum.